How Does Feedback Impact Training in Audio-Visual Speech Perception?

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation *with distinction* in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

By

Amy Ranta

The Ohio State University
December 2010

Project Advisor: Dr Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Integration of visual and auditory speech cues is a process used by listeners in compromised listening situations, as well as in normal environments, as exemplified by the McGurk effect (McGurk and McDonald, 1976). Audio-visual integration of speech appears to be a skill independent of the ability to process auditory or visual speech cues alone. Grant and Seitz (1998) argued for independence of this process based on their findings that integration abilities could not be predicted from auditory-only or visual-only performance. Gariety (2009) and James (2009) further supported this argument by training listeners in the auditory-only modality with degraded speech syllables, then testing those listeners in the auditory-only, visual-only, and audio-visual conditions. Their results showed an increase in auditory-only performance, but no improvement in integration. Recently, DiStefano (2010) conducted a training study in which listeners were trained in the audio-visual modality with degraded speech syllables. Results showed that the performance increased only for the audio-visual conditions, and did not increase in the auditory-only or visual-only conditions. Interestingly, performance improved only for stimulus pairs that were "congruent" (i.e. the auditory and visual inputs were the same syllable) and did not increase for "discrepant" stimuli (i.e. the auditory and visual inputs were different syllables).

It is possible that the feedback provided in DiStefano's study impacted their pattern of results. However, the question remains as to whether integration of discrepant stimuli can be trained. In the present study, five listeners received ten hours of training sessions in the audio-visual condition with degraded speech signals similar to those used by Shannon et al. (1995). The feedback given during training was designed to encourage McGurk-type combination and fusion responses, in contrast to DiStefano's study, in which feedback was given to encourage

responses that matched the auditory signal. A comparison of pre-training and post-training scores showed little to no improvement in auditory-only performance, and a slight decrease in visual-only performance for congruent stimuli. Further, a substantial increase in McGurk-type responses was seen from pre- to post-test for discrepant stimuli. These results provide further support that integration is an independent process, and that the feedback provided strongly influences response patterns. This strong lack of generalization from training also should be incorporated into designing effective integration training programs for aural rehabilitation.

Acknowledgments

I would like to thank my advisor, Dr. Janet M. Weisenberger, for providing me with the opportunity to work with her on this thesis. Through her guidance, patience, and support I was able to gain a valuable experience and grow academically and professionally. I would also like to thank my family and friends for their support and encouragement, and my subjects for their time and flexibility.

This project was supported by a Division of Social and Behavioral Sciences Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

Although it is generally thought that people rely primarily on the information contained in an auditory signal to understand speech, it is not the only process we use. It has long been known that people use information from both auditory and visual inputs to understand speech, especially when the auditory signal is being compromised (i.e., noisy environment, hearing loss, etc.). However, the McGurk effect demonstrates that even people with perfect hearing, listening to an uncompromised auditory signal, use inputs from both the auditory and visual modality to perceive and understand speech.

Research by McGurk and MacDonald has shown that visual cues are used even when the auditory signal is not compromised (McGurk & MacDonald, 1976). In their study, listeners were presented with discrepant auditory and visual stimuli to observe the amount of audio-visual integration. When presented simultaneously with the visual stimulus /ba/ and the auditory stimulus /ga/, most subjects reported that they perceived the syllable /da/, a "fusion" response, so-called because listeners appear to fuse the places of articulation of the two inputs to produce an entirely different response. When the auditory and visual stimuli were reversed (i.e., a visual /ba/ and a auditory /ga/), the most commonly reported response was /bga/, a "combination" response which occurs when the visual place of articulation is so salient that a fused syllable cannot be formed. This audio-visual integration phenomenon, which occurs when the brain cannot ignore strong input from either modality, became known as the McGurk effect. It is well known today that the occurrence of audio-visual integration is a behavior that is both automatic and unconscious for listeners. To understand audiovisual speech integration more completely, it is important to evaluate the auditory and visual information that is available to the perceiver in a speech utterance.

Auditory Cues for Speech Perception

In most cases, the auditory signal alone contains enough information for the listener to accurately identify speech sounds. There are three main cues in the auditory signal for identifying consonants, which include place of articulation, manner of articulation, and voicing. Place of articulation refers to the physical place in the mouth where the sound is produced, or the location where the airstream is obstructed in the vocal tract. These locations include bilabials, labiodentals, interdentals, alveolars, palatal-alveolars, palatals, and velars. Manner of articulation refers to the physical orientation of the articulators during sound production. Manner includes stops, fricatives, affricates, liquids, and glides. Voicing refers to whether or not vibration is present in the vocal folds during sound production. Vocal fold vibration during speech production indicates a voiced sound, whereas sounds without vibration are voiceless. This information regarding the auditory signal can be found in the spectral and temporal envelopes of a speech waveform (Ladefoged, 2006).

Information contained in vowels is most likely conveyed in the overtone structure, or formant, of the vowel. Formants are groups of overtones corresponding to a resonating frequency of the air in the vocal tract. Each vowel sound has a unique formant structure, which can be characterized by three formants (F1, F2, F3). Vowels can be largely distinguished from the first two formants, F1 and F2. F2 is the higher of the two formants, and slopes downward in frequency during the production of most vowels. F1 is lower than F2, and slopes upward for some vowels /i, I, e, E/, and downward for other vowels /ae, a, ɔ, o/. Vowels also contain information in their third formant, but not as much as in F1 and F2 (Ladefoged, 2006).

Visual Cues for Speech Perception

Although much of the information needed to identify a speech sound is transmitted through the auditory signal, McGurk and MacDonald helped to show the significance of the visual input in speech perception, regardless of whether the auditory information is compromised. Though the information in visual cues is important, it does not contain nearly as much information about the characteristics of the speech sound as the auditory signal. Place of articulation is the only characteristic that can be reliably observed, and even this information is often ambiguous (Jackson, 1988). Since there is no information concerning voicing in the visual signal, and limited information on the manner of articulation, it is sometimes impossible to identify a sound by visual information alone.

The obstacle in identifying speech sounds from visual information alone occurs because there are groups of phonemes, such as /p, b, m/, that are visually indistinguishable. These groups of visually identical phonemes are called visemes (Jackson, 1988). Visemes, or visual phonemes, differ in manner of articulation and voicing characteristics, but have the same place of articulation. Jackson showed that talkers who created more viseme groups were easier to speechread than talkers who created fewer. Talkers have individual characteristics that can affect their level of intelligibility, such as head and eye movements, gestures, and even facial hair. Visual cues to speech can be especially useful in situations where the auditory signal is compromised.

Speech Perception with Reduced Auditory and Visual Signals

Research has shown that speech can still be highly intelligible even in situations where the signal is compromised, partly because there is more information than necessary in the acoustic speech signal to identify the phoneme. In a study by Shannon et al. (1995), speech sounds were degraded by removing the fine structure information and replacing it with band-limited noise, all while preserving the temporal envelope. This process of degrading speech sounds is similar to the signal produced by cochlear implants. Shannon found that listeners were still able to indentify speech sounds with as little as 3 to 4 channels of sound. This study shows that much of the information in the auditory signal is redundant and not absolutely necessary for speech recognition. This study was expanded in 1998 by Shannon and his colleagues into four experiments, including varying the location of band division, warping the spectral distribution of envelopes, shifting the frequencies of envelope cues, and spectral smearing. Findings showed that warping the spectral distribution and shifting the tonotopic organization of the envelopes had the most negative effect on intelligibility, whereas the other manipulations showed that exact frequency cutoffs and overlapping of the bands do not have much effect on speech intelligibility (Shannon et al., 1998).

Another study that examined how reduced auditory input can impact speech perception was performed by Remez et al. (1981). In this study, the auditory signal was reduced to three sine waves that represented the formants of the original speech sound. Using this method, known as sine wave speech, Remez and his colleagues still found that speech was highly intelligible (Remez et al., 1981). This produced further evidence that speech can still be intelligible even with a highly degraded acoustic signal.

While studies by Shannon and Remez showed that the auditory information does not need to be perfect to aid in intelligibility, Munhall et al. (2004) conducted a study to observe the effects of a compromised visual signal on speech perception abilities. In this study, visual images that had been degraded through band-pass and low- pass spatial filtering were presented along with auditory signals in noise. Results showed that subjects had highest levels of speech intelligibility in the mid-range filter band with a center spectral frequency of 11/cycles per face. These results also indicate that high spatial frequency information is not needed for speech perception. Munhall et al. concluded that like compromised auditory cues, visual images do not need to be perfect to aid in speech perception (Munhall et al., 2004).

Audio-Visual Integration of Reduced Info Stimuli

Studying audio-visual integration with degraded auditory stimuli can provide an approximation to the situation encountered by hearing impaired persons, and thus is useful for designing aural rehabilitation programs, because it provides data on how visual inputs are used to complement a reduced auditory signal. A previous study in our laboratory by Feleppelle (2008) examined the auditory signal's role in audio-visual integration. Her experiment was conducted to examine whether reducing the information in the auditory signal was a contributing factor to the variability in audio-visual integration benefits across listeners. Listeners' speech perception abilities were tested in auditory-only, visual-only, and audio-visual conditions at four different levels of auditory degradation. The stimuli were degraded using 2, 4, 6, and 8 bandpass filter channels in a manner similar to that used by Shannon et al. (1998). Results showed that listeners were able to achieve high levels of integration from audio and visual cues even with a highly degraded auditory signal. However, even with the highest levels of auditory signal degradation,

the amount of audio-visual, defined as the difference between the audio-visual performance and the best single modality, did not change.

In a study by Grant and Seitz (1998), audio-visual integration of hearing impaired subjects was evaluated. In this case, the auditory signals were congruent and discrepant nonsense syllables degraded due to participants' hearing loss. The congruent stimuli presented the auditory cue synchronized with the visual cue, whereas discrepant stimuli have the auditory and visual cues "out of sync" either through timing differences or by dubbing an auditory signal with a different visual cue. The subjects were presented with the syllables in three conditions: auditory (A), visual (V), and audio-visual (AV). Results indicated that even when audio input is poor, listeners can increase their speech perception with added visual information. Because the amount of audio-visual integration could not be predicted from auditory-only or visual-only performance, Grant suggested that integration is a cognitive skill independent of auditory or visual processing.

Effects of Training in Recent Studies

Recent studies in our laboratory have examined the effects of training in different modalities on audio-visual integration abilities. James (2009) and Gariety (2009) studied whether providing auditory training with degraded speech would improve integration performance. Their studies employed two types of speech distortion, one similar to that used by Shannon, and one similar to that used by Remez in which speech syllables were reduced to a few sine waves that followed the formant structure. Participants were trained under auditory-only conditions. Results of both studies showed that training in the auditory modality improved auditory performance, but

not audio-visual integration. Their results provide further evidence that integration is a skill that is independent of auditory and visual abilities alone.

Very recently, a study by DiStefano (2010) showed preliminary evidence that training in the audio-visual modality for degraded speech signals improves integration efficiency without improving audio or visual performance alone, again supporting the idea of independent processing. She trained participants with two types of syllables: congruent, in which the auditory and visual components of the speech were the same syllable, and discrepant, in which the auditory and visual components were different syllables, like those used by McGurk and MacDonald. She observed an increase in audio-visual integration for the congruent stimuli after training. However, for stimuli constructed to elicit McGurk-type integration, an increase in McGurk-type response was not observed. This lack of integration may be attributable to the specific feedback provided to listeners during training. In her study, listeners were given feedback reinforcing the auditory information that was presented, as opposed to the visual information, or the McGurk-type combination or fusion responses.

These results are of particular interest because they suggest that training is highly specific to particular stimuli and modalities and does not generalize. If true, then these results have important implications for the design of aural rehabilitation training for hearing-impaired persons, indicating the use of a much broader variety of talkers and speech types than are frequently observed in aural rehabilitation programs.

Present Study

Though there is now evidence that audio-visual integration is a skill that can be trained, the question remains as to how feedback during training will affect McGurk-type integration

responses. The present study investigated whether McGurk-type integration can be trained. Five normal hearing listeners each received 10 hours of training similar to that in DiStefano's study, but feedback provided during training was designed to specifically elicit McGurk-type integration responses. It was hypothesized that McGurk-type fusion and combination responses would increase after training sessions were completed, as well as all integration performance. It was also expected that auditory-only and visual-only performance would not show improvement from integration training. The results from this study may provide further insight into the process of audio-visual speech integration and improve methods used in aural rehabilitation.

Chapter 2: Method

Participants

Participants in this study included 5 listeners. Three females and two males, ages 21-22 years, participated. All 5 reported having normal hearing as well as normal to corrected vision. Four subjects had no background in Speech and Hearing Science, and though one subject had a background, she reported no previous knowledge of the McGurk effect. Participants were compensated $160 for their time. Materials previously recorded by five talkers, 2 male, 3 females, were used as stimuli.

Stimuli Selection

A limited set of eight syllables were presented for the study. All syllables presented to listeners had the same conditions:

1. Pairs of stimuli were minimal pairs, meaning they only differed in initial the consonant.

2. The vowel /ae/ was used for all stimuli because it does not involve lip rounding or lip extension, which could make speech reading more difficult.

3. Each category of articulation, including place (bilabial, alveolar, velar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced), was represented in the stimulus syllables.

4. All were presented individually and without a carrier phrase.

Stimuli

For each condition the same set of single-syllable stimuli was used:

Bilabial: bat, mat, pat

Alveolar: sat, tat, zat

Velar: cat, gat

The following dual-syllable (dubbed) stimuli were used in audio-visual conditions. The first syllable represents the visual stimulus and the second syllable the auditory stimulus.

bat-gat

gat-bat

pat-cat

cat-pat

Stimuli Recording and Editing

Stimuli from resent studies in our lab (e.g. James, 2009 and DiStefano, 2010) were used in this study to permit comparable results. Speech samples from five talkers, two male and three female, were degraded using a MATLAB script designed by Delgutte (2003). The speech signal was filtered into two broad spectral bands. Then the fine structure of each band was replaced with band limited noise, while keeping the temporal envelope intact. The result was a 2-channel stimulus similar to those used by Shannon et al. (1998). The degraded auditory stimuli were then dubbed onto the visual stimuli using Video Explosion Deluxe, a commercial video editing program.

Finally, the software program Sonic MY DVD was used to burn the stimulus sets onto DVDs. Four DVDs were created for each of the five talkers, each with sixty stimuli in random order to eliminate the possibility of memorization from the participants.

Visual Presentation

Each participant was pre-tested under degraded auditory (A), visual (V), and audio-visual (A+V) conditions with no feedback, followed by training with degraded audio-visual presentation with feedback. For presentation of visual portion of the stimulus, a 50 cm video monitor was placed approximately 60 cm outside of the window of a sound attenuating booth. The monitor was darkened and positioned at eye level of participants, and about 120 cm away from where they were seated inside the booth.

Degraded Auditory Presentation

The degraded auditory stimuli were presented from the headphone output of the DVD player through 300-ohm TDH-39 headphones at a level of approximately 75 dB SPL.

Testing Procedure

Testing was conducted in the Ohio State University's Speech and Hearing Department. Participants were instructed to read over a set of instructions explaining the procedure and listing a closed-set of 14 response possibilities. The response set included more response possibilities than the 8 stimuli actually presented to include options that might reflect McGurk-type fusion or combination responses for the discrepant stimuli. The additional possibilities included syllables dat, nat, pcat, ptat, bgat, and bdat.

Participants were individually tested in a sound -attenuating booth facing a video monitor placed outside the booth. Auditory stimuli were transmitted through headphones inside. The examiner recorded and scored the participant's verbal responses as heard through an intercom system. Each participant was first administered a pre-test using stimuli selected from a set of 15 DVDs, three DVDs for each of the five talkers, each containing 60 randomly ordered syllables. In the pre-test, the listeners were presented with one DVD from each talker in each of the three conditions (A, V, and A+V). Each DVD in the audio-visual condition contained 30 stimuli that were congruent and 30 stimuli that were discrepant, intended to elicit McGurk-type responses. Participants were instructed to listen to/watch each DVD and to verbally respond what they perceived for each syllable. No feedback was provided during the pre-test.

The pre-test was followed by five A+V training sessions, each about 45 minutes to an hour in length. Each session included presentation of one DVD from each of the five talkers. Trial-by-trial feedback was provided to the participants. For congruent stimuli, if the participant responded with an incorrect response, the examiner would verbally provide the correct answer through the intercom. If the participant provided the correct answer then the examiner would visually reinforce the participant with a head nod. For discrepant stimuli, the appropriate McGurk-type component feedback was given to participants as the correct response. The feedback was given as follows, with the first syllable representing the visual stimulus, second syllable representing the auditory, and the third representing the McGurk-type feedback provided:

bat-gat bgat

gat-bat dat

pat-cat	pcat
cat-pat	tat

The choice to provide the McGurk type component feedback was made to determine the effects of training in the A+V condition with discrepant McGurk-type stimuli, as well as congruent stimuli, throughout the study.

A mid-test, using the same procedure as the pre-test, was administered following the first five training sessions. No feedback was provided. Five more auditory plus visual training sessions, with the same procedure as the first five, followed the mid-test. Finally, a post-test was conducted, again with no feedback. Each test took approximately 2-3 hours, and the 10 training sessions took approximately 8-10 hours. Training sessions were broken up into 1 to 2 sessions at a time, and participants were encouraged to take frequent breaks to prevent fatigue.

Data were analyzed by comparing mean performance of arcsine-transformed percentages. Repeated measures analyses of variance were used to examine main effects and interactions of the variables test (pre, mid, post) and modality (A, V, A+V). Means comparisons pinpointed the specific nature of any significant effects. All tests were performed with α=.05.

Chapter 3: Results and Discussion

Results of the pre-test, mid-test, and post-test were analyzed to determine how specific training affected identification performance in the audio-visual condition with degraded stimuli. Results are presented first for percent correct performance with congruent stimuli, followed by results for discrepant stimuli.

Percent Correct Performance

Figure 1 shows the overall percent correct performances for congruent stimuli, for the auditory-only (A), visual-only (V), and audio-visual (A+V) conditions for the pre-test, mid-test, and post-test, averaged across talkers and listeners. An increase was seen in audio-visual performance from pre- to post-test, while auditory-only performance did not show improvement and remained relatively constant and visual-only performance actually decreased from pre- to post-test. A two-factor repeated measures analysis of variance (ANOVA) was performed on arcsine-transformed percentages to assess the significance of their changes in performance. ANOVA results indicated a significant main effect of test (pre, mid, post), $F(2,8)=94.68$, $p<.001$. A significant main effect of modality (A, V, A+V) was also observed, $F(2,8)=16.37$, $p=.003$. Finally, a significant test x modality interaction was found, $F(4,16)=14.34$, $p=.001$. Pairwise comparisons indicated significant differences in pre-mid ($p=.004$), mid-post ($p=.001$), and pre-post ($p<.001$) means, as well as in all modalities ($p\leq.001$) means. The increase in A+V performance, as well as the decrease in visual-only performance across tests, is consistent to DiStefano's (2010) findings in direction, although DiStefano observed greater improvement in A+V from pre-test to post-test. DiStefano's results are shown in Figure 2. This agreement between the two studies supports the idea that integration is a process independent of auditory

and visual conditions, given that training in the audio-visual condition did not result in improvement in auditory-only or visual-only performance.

Figure 3 indicates differences in performance for stimuli produced by different talkers. This figure shows auditory-only, visual-only, and audio-visual pre-test responses averaged across listeners, for each talker. All of the five talkers were most intelligible in the audio-visual condition. The visual-only condition was superior to the auditory-only for three of the five talkers, while auditory-only was perceived better than visual-only for the other two talkers.

Figure 4 shows auditory-only, visual-only, and audio-visual mid-test responses averaged across listeners, for each talker. After the initial five training sessions, all five talkers were perceived better in the auditory-only condition than in the visual-only condition, and correct responses for the visual-only condition actually decreased. Intelligibility in the audio-visual condition remained relatively constant for all talkers compared to the pre-test results. This is a somewhat surprising difference from the results of DiStefano's study, in which four out of five talkers showed substantial improvement in audio-visual intelligibility from pre-test to mid-test, and raised some question about the efficacy of training across the first five sessions.

Figure 5 shows auditory-only, visual-only, and audio-visual post-test responses averaged across listeners, for each talker. Again, for all five talkers, auditory-only intelligibility was better than visual-only. There was a slight improvement in audio-visual intelligibility from the mid- to post-test. This suggests that the full ten hours of training continued to produce improvements in the audio-visual condition. Also, the slight increase in auditory-only intelligibility across tests shows that there may have been some learning in the auditory-only condition from pre-test to post-test, though the increase was quite small. The difference in response patterns of mid- to

post-tests from DiStefano's study to this study could be due to the specific McGurk-type feedback provided in each study. Mainly, the decrease in visual-type responses across tests was a result of the larger number of McGurk-type responses provided by the listener, even though the stimuli in this case were congruent.

Figure 6 shows the amount of audio-visual integration, where integration is defined as the difference between audio-visual performance and the better single modality, auditory or visual, averaged across listeners, for each talker. In the present study, while integration performance increased with all talkers from pre to post-test, the amount of improvement across talkers was very different. Talkers JK and KS showed the greatest increase in integration from pre-to post-test, with integration of 15-16% in the pre-test and 29-33% in the post-test. It is important to note that the large improvement could have been due to talkers JK and KS producing the lowest level of integration from the beginning, providing much opportunity for improvement. Talkers DA, EA and LG all showed similar amounts of improvement in integration from pre to post-test, with increases ranging from 6-8%. Talkers EA and LG began with high levels of integration in the pre-test, so it is possible that they did not have much room to improve. Talker DA had the median integration level of 22% in the pre-test, and only had 28% in the post-test. This talker is interesting because he produced a medium level of integration compared to other talkers, and with only a slight increase in integration from pre-test to post-test, he ended as the talker producing the least amount of integration. DiStefano's amount of integration by talker results are shown in Figure 7. These results showed improvement that did not vary so markedly across talkers. The present study showed a similar range of integration in the pre-test, and slightly higher levels in the post-test than DiStefano's. These differences in the results between studies can be attributed to differences between listeners.

Figure 8 shows the amount of audio-visual integration for individual listeners. Four out of five listeners showed improvements in integration, with variability ranging from 8-18%. The varied range of integration improvement could simply be due to listener differences. One listener, TM, actually showed a decrease of 9% in the amount of integration from pre to post-test. This could be due to an increase in McGurk–type combination and fusion responses for the congruent stimuli in the post-test, as a result of specific feedback that encouraged these responses during training. It should also be noted that listener TM showed the highest level of integration of all listeners in the pre-test at 44%, which was 22% higher than any other listener. DiStefano's results for integration for individual listeners are shown in Figure 9. Results from DiStefano's study showed an increase in integration for all listeners from pre to post-test, but the increase for three of the five listeners was much smaller than any of the four improved listeners in the present study, and two of her listeners showed much larger integration from pre to post-test. The large range of listener integration improvement in DiStafano's study was not shown in the listeners of the present study. This could be attributable to the time each individual listener took to complete the training and testing. DiStefano's listeners were tested at their convenience and in some cases took much time off between training sessions and testing, whereas the listeners in the present study were instructed to schedule their training and testing sessions to complete within a timeframe of two to three weeks. The better performers in DiStefano's study were those who completed training within the two-week timeframe.

Integration of Discrepant Stimuli

In addition to congruent stimuli presentations, listeners were also presented with discrepant stimuli, in which the auditory stimulus differed from the visual stimulus. While there is no "correct" response for these stimuli, feedback was given during training to encourage McGurk-type combination and fusion responses. The responses were categorized according to whether the listener chose a response which matched the auditory or visual stimulus, or chose some "other" response which matched neither the auditory or visual stimulus. The "other" responses were then categorized as either a fusion response, combination response, or neither.

Figure 10 shows the overall responses for discrepant stimuli for all tests, averaged across talkers and listeners. Across all tests, the percentage of responses categorized as auditory was extremely low, 2% or less, likely due to the degree of degradation of the auditory signal. In the pre-test, listeners relied heavily on visual information, and a high percentage of their responses, 61%, matched visual stimuli. Responses categorized as "other" were much lower, at about 37%. DiStefano's results are shown in Figure 11. Interestingly, in DiStefano's study, a much lower percentage of "other" responses, 15% was observed, which could simply be due to differences across listeners. By the mid-test, the percentage of visual responses in the present study dropped substantially, with the percentage of "other" responses rising to 84%. The post-test percentages were similar to those of the mid-test for auditory, visual, and "other" responses. This suggests that training listeners to produce responses that differ from the auditory or visual stimulus actually presented can be effective as quickly as the first five hours of training. A two-factor, repeated measures ANOVA for the percent response data revealed no significant main effect of test, $F(2,4)=.286$, ns, but did show a significant main effect of response type (auditory, visual, other), $F(2,4)=93.72$, $p<.001$. A significant test x response type interaction was also evident,

F(4,16)=95.9, p<.001.These findings are very different from DiStefano's study, where listeners relied on the visual modality across pre-, mid-, and post-tests. This suggests that the specific type of feedback given to listeners can substantially influence their responses.

Figure 12 further analyzes "other" responses from Figure 10, and shows percent McGurk-type integration for discrepant stimuli, averaged across talkers and listeners. In the pre-test, there were a relatively low number of fusion or combination responses. This lack of McGurk-type responding could be attributed to the fact that combination responses such as "pcat" or "bgat" are not permissible phoneme sequences in English, and thus listeners were not used to hearing or producing these responses. By the mid-test, there was a distinct increase in McGurk-type responses from 24% in the pre-test to 79% in the mid-test. The fusion responses increased from about 16% in the pre-test to 38% in the post-test. However, an even greater increase was observed in the combination type responses, moving from 8% in the pre-test to 41% in the mid-test. From the mid- to post-test, an additional increase was observed in McGurk-type responses. Means comparisons for these results showed a significant increase in fusion responses from pre-test to post-test, t(4)=12.9, p<.001, as well as a significant increase in combination responses, t(4)=16.99, p<.001. Correspondingly, a significant decrease in neither responses was found, t(4)=29.82, p<.001. This strong increase in McGurk-type responses across training suggests that the feedback provided to listeners greatly impacts their response types. Even more importantly, the effects of feedback are very specific and do not appear to generalize.

The pattern of results in the present study was rather different from that observed by DiStefano (2010), whose results are shown in Figure 13. In DiStefano's study, there was little increase in McGurk-type responses seen across training, whereas the present study shows large increases in these types of responses. Further, the present study, there was a relatively even

distribution of combination and fusion type responses reported by listeners, whereas in DiStefano's study there was only a small increase in fusion responses. Again, this was likely attributable to the differences in feedback during training in the two studies.

Chapter 4: Summary and Conclusions

Results of testing indicate that training in the audio-visual condition does result in a significant improvement in audio-visual integration performance; however, the specific feedback provided can significantly affect listeners' response patterns. Four out of five listeners in the present study showed improved integration performance, while one listener actually showed a decrease in integration performance from pre to post-test. This decrease was interestingly due to the increase in McGurk-type responses from that listener for congruent stimuli in the post-test that were not present in the pre-test. While performance in the audio-visual condition improved, performance in the auditory-only condition did not show much improvement, and the visual-only condition performance actual decreased, again as a result of the increase in McGurk-type responses. These results provide additional support for Grant and Seitz's (1998) argument that integration is a skill that is separate from processing in either individual modality.

Since an increase in performance was only seen in the modality in which listeners were trained, it might be assumed that training in a specific modality is limited to that modality and does not generalize. However, the increase of McGurk-type responses seen in the post-test for congruent stimuli suggests that some training can generalize too much, and affect overall performance. Future studies should investigate generalization of integration across talkers, perhaps by training listeners with some talkers and then testing listeners with other talkers.

It is important to understand the type of stimuli that best improve integration skills in order to design the best aural rehabilitation program in training hearing impaired persons to make use of residual hearing and other sensory cues. Based on the lack of generalization of results from training across modalities, it may be that the best aural rehabilitation program would include training with a range of talkers as well as a range of types of stimuli used. The decrease

in visual-only performance across tests in the present study further supports the idea that the most successful aural rehabilitation program would include training in the visual, auditory, and audio-visual conditions. For a specific listener in need of an aural rehabilitation program, training in all three conditions should be specialized to individual specific needs and skills to maximize the possible benefits.

Chapter 5: References

DiStefano, S. (2010). *Can audio-visual integration improve with training?* Senior Honors Thesis, The Ohio State University.

Feleppelle, N.M. (2008). The *role of the auditory signal in auditory-visual integration.* Audiology Capstone Project, The Ohio State University.

Gariety, M. (2009). *Effects of training on intelligibility and integration of sine-wave speech.* Senior Honors Thesis, The Ohio State University.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104 (4),* 2438-2450.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90(5),* 99-114.

James, K. (2009). *The effects of training on intelligibility of reduced information speech stimuli*. Senior Honors Thesis, The Ohio State University.

Ladefoged, P. (2006). A *Course in Phonetics-Fifth Edition.* Boston : Wadsworth.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

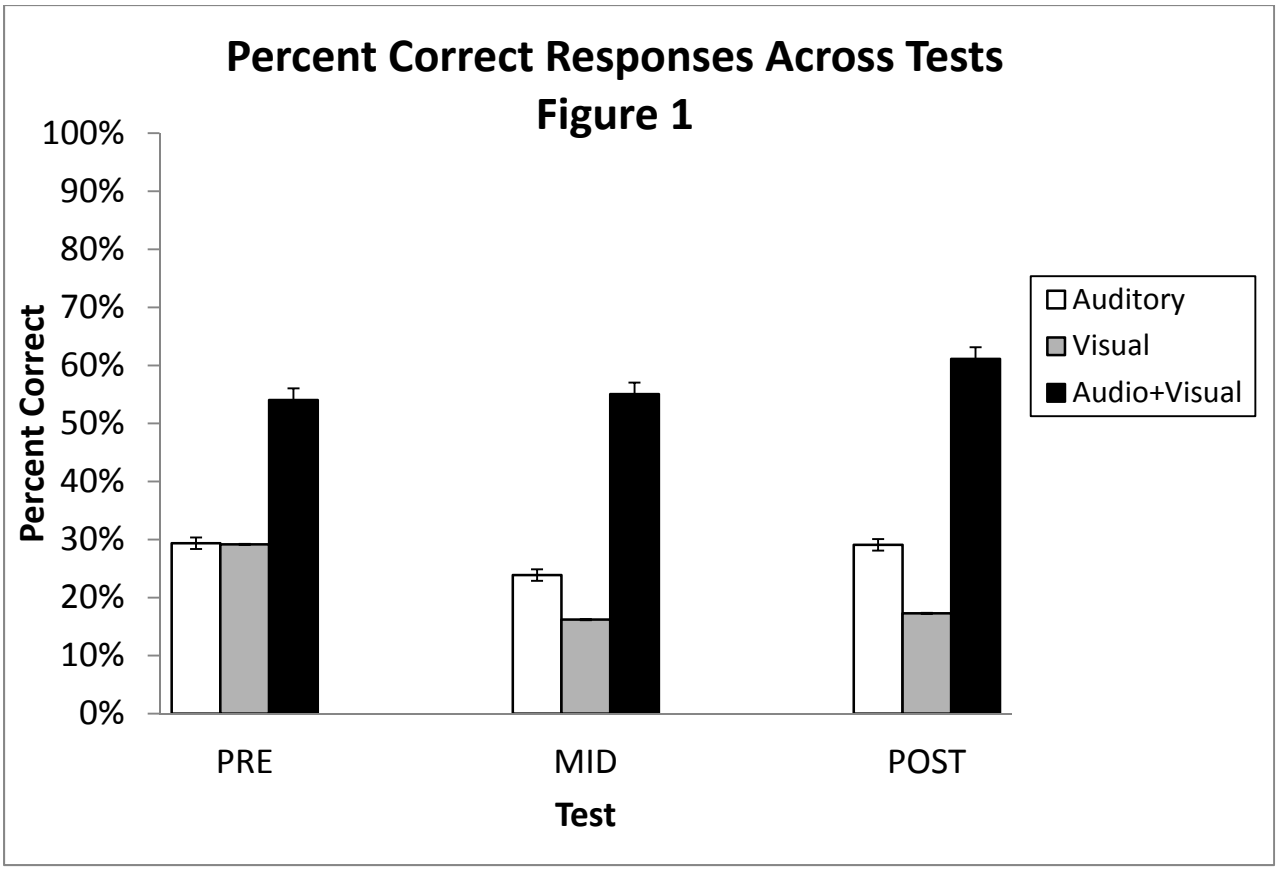
Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics, 66,* 574-583.

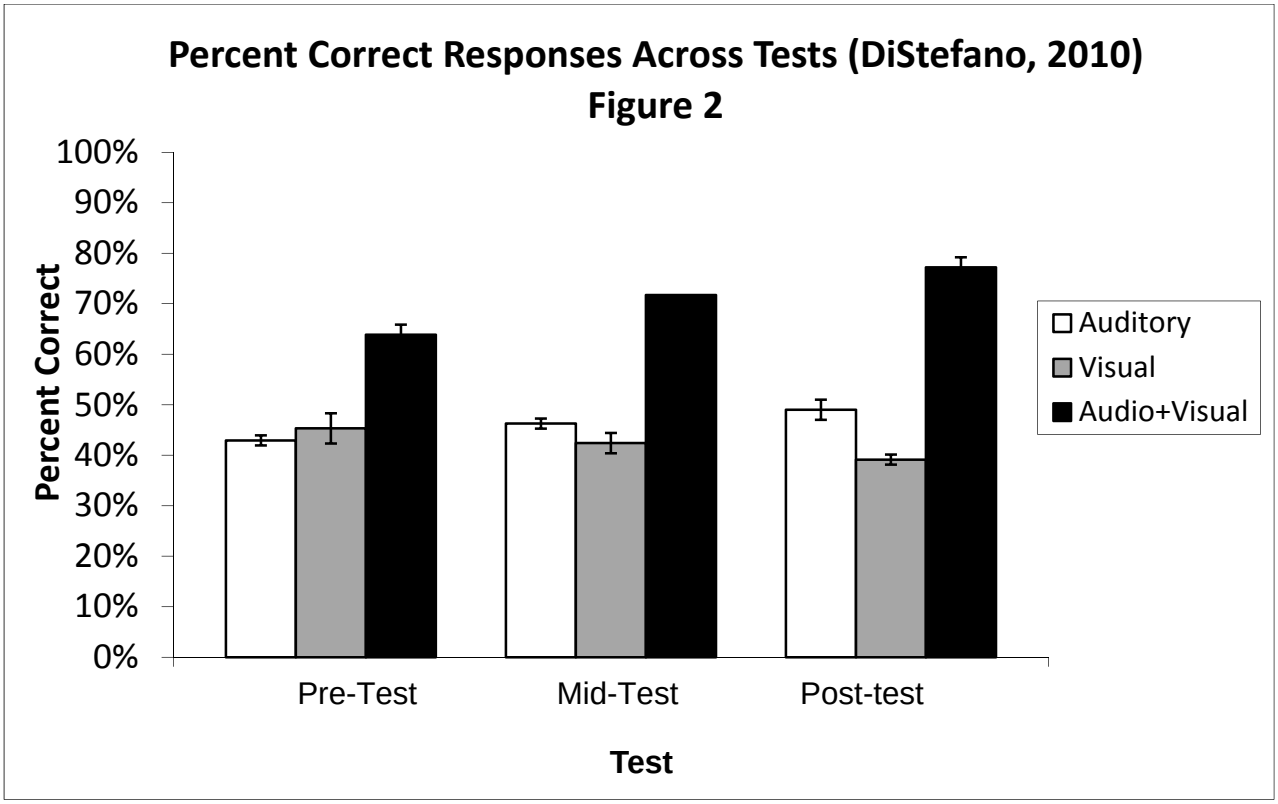
Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional cues. *Science, 212 (4497),* 947-950.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

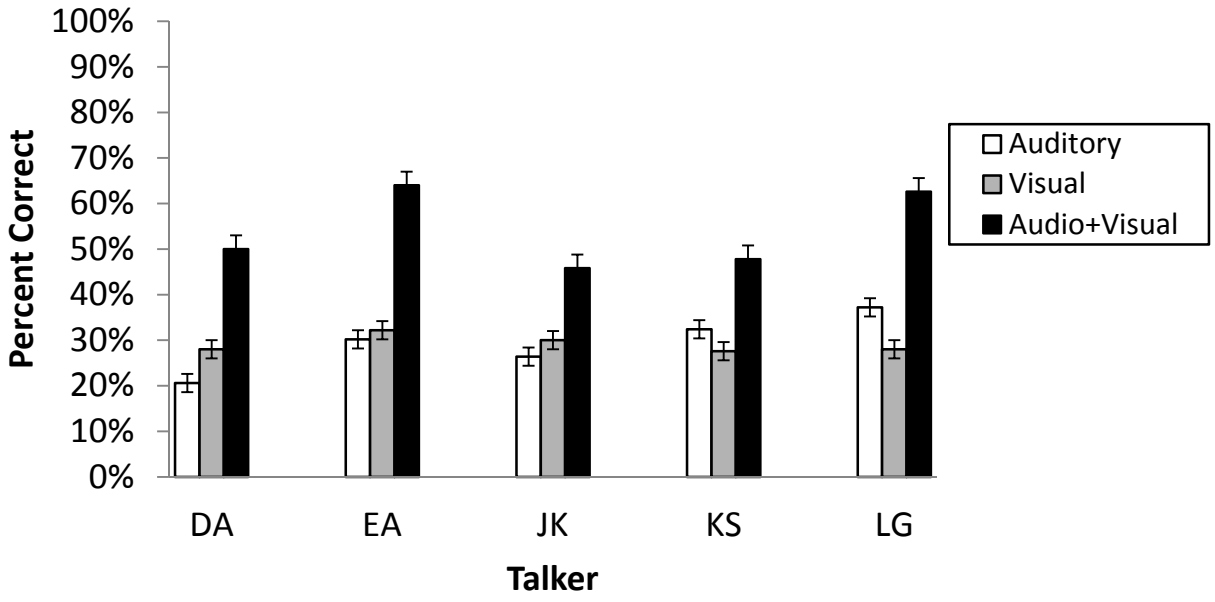
Shannon, R.V., F.G., Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America*, *104 (4),* 2467-2475.

List of Figures

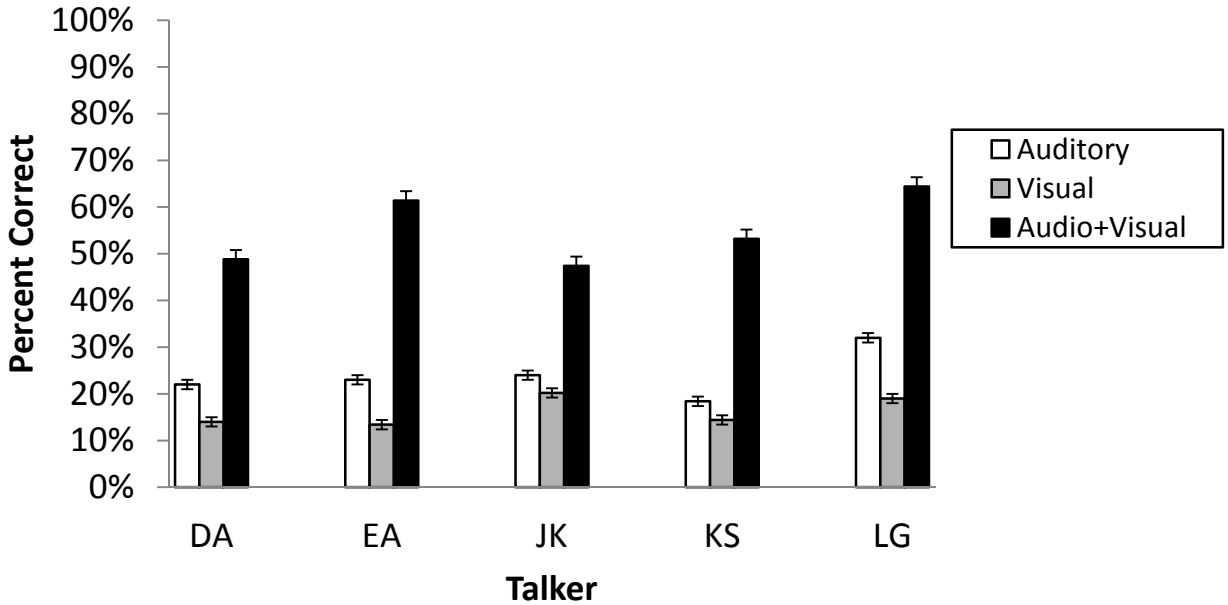

Percent Correct Responses Across Tests
Figure 1
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent Correct
Auditory
Visual
Audio+Visual
PRE
MID
POST
Test

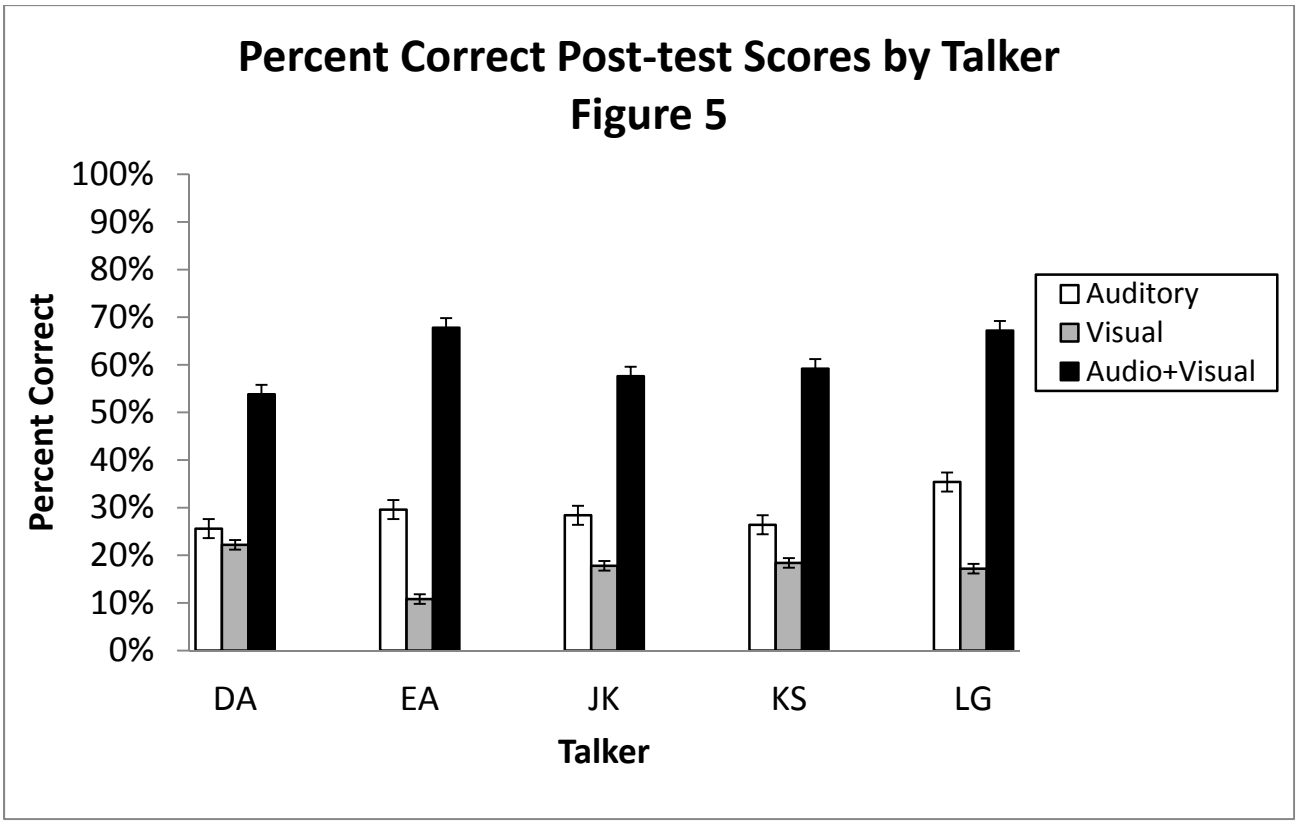

Percent Correct Responses Across Tests (DiStefano, 2010)
Figure 2
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent Correct
Auditory
Visual
Audio+Visual
Pre-Test
Mid-Test
Post-test
Test

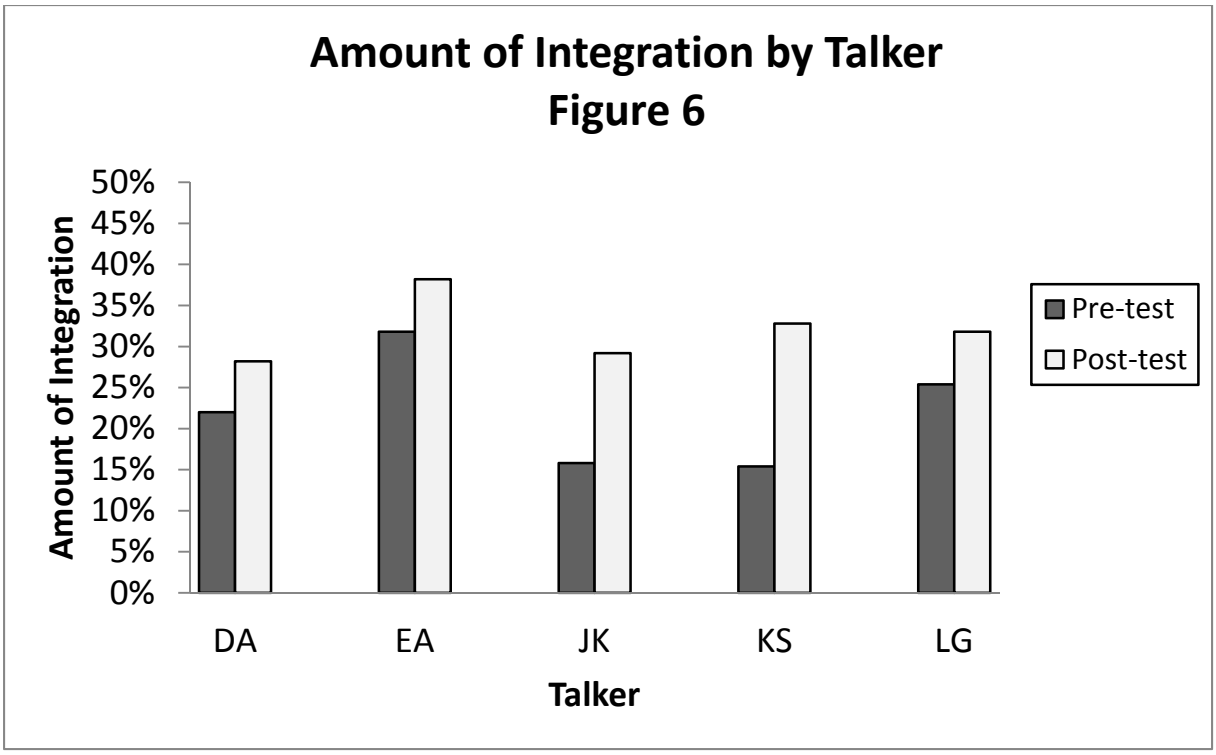
Percent Correct Pre-test Scores by Talker
Figure 3

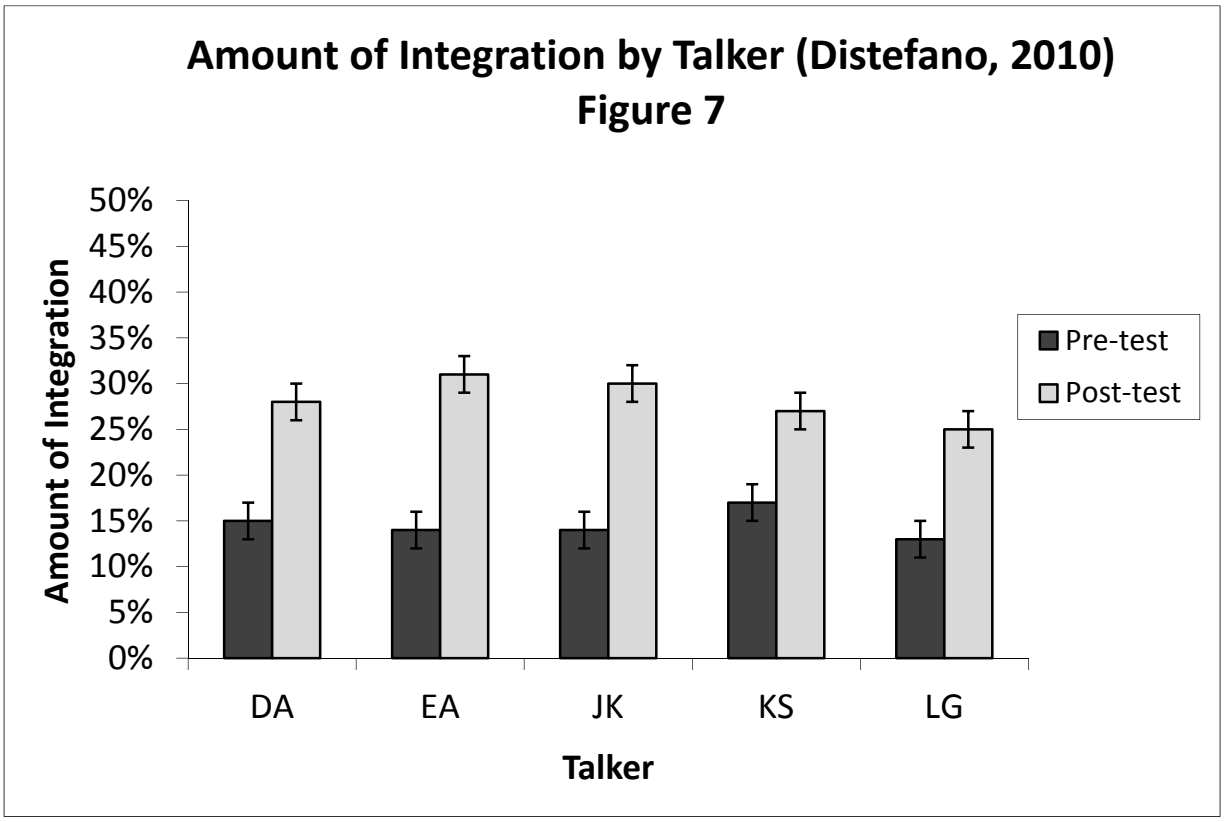


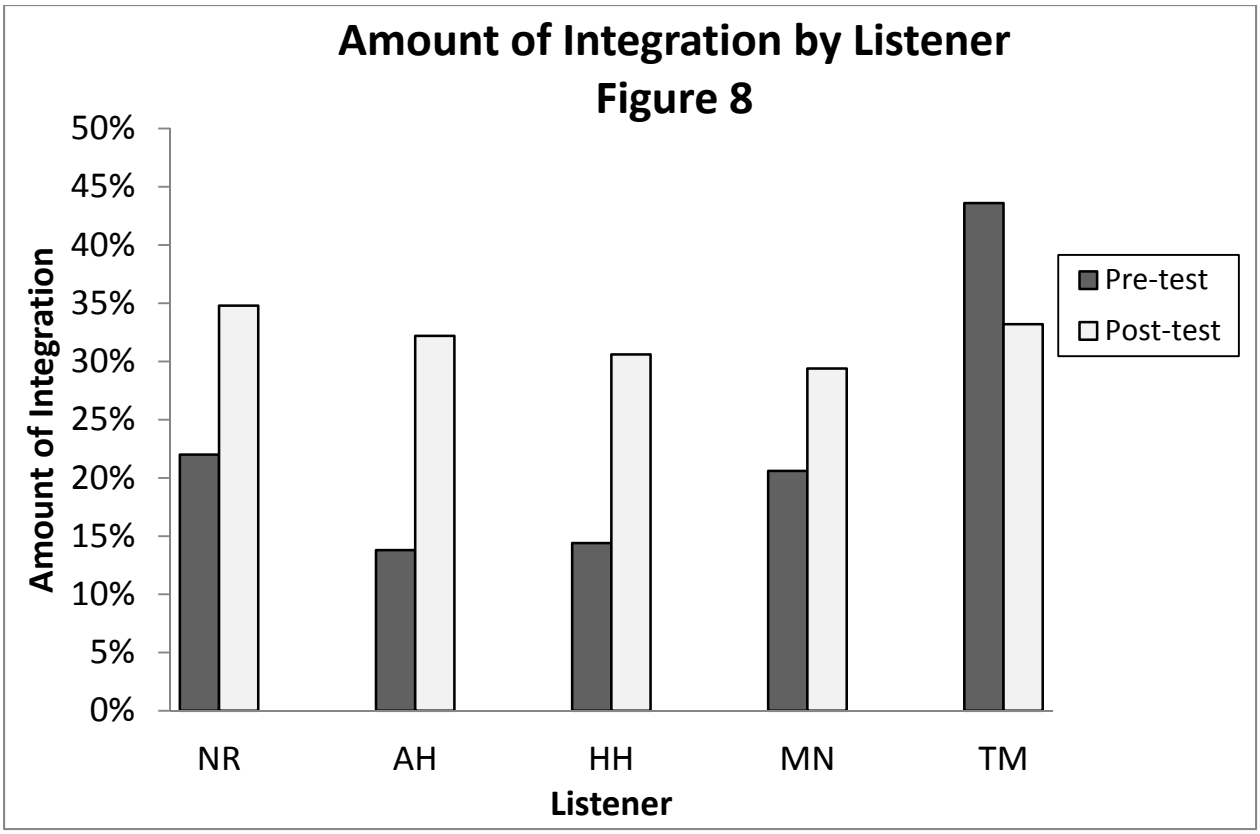
Percent Correct Mid-test Scores by Talker
Figure 4

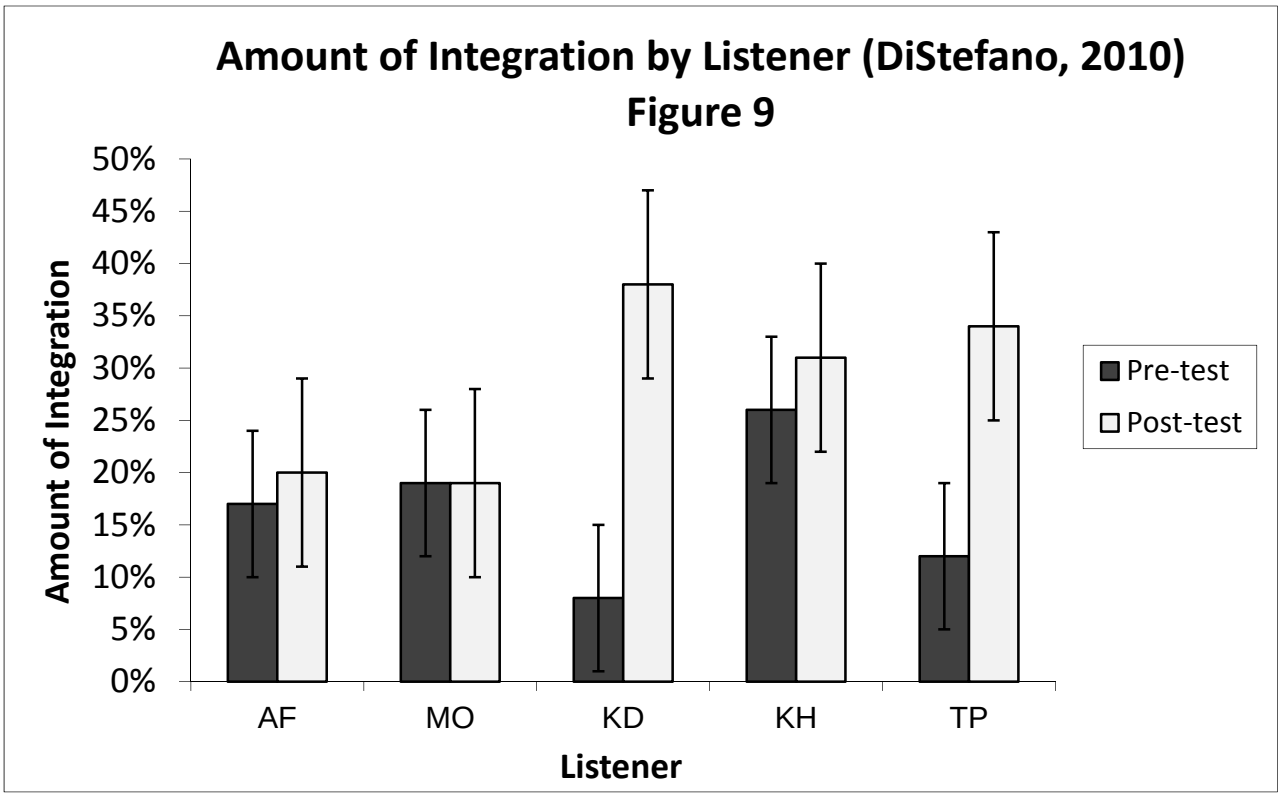



Percent Correct Post-test Scores by Talker
Figure 5
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent Correct
DA
EA
JK
KS
LG
Talker
Auditory
Visual
Audio+Visual

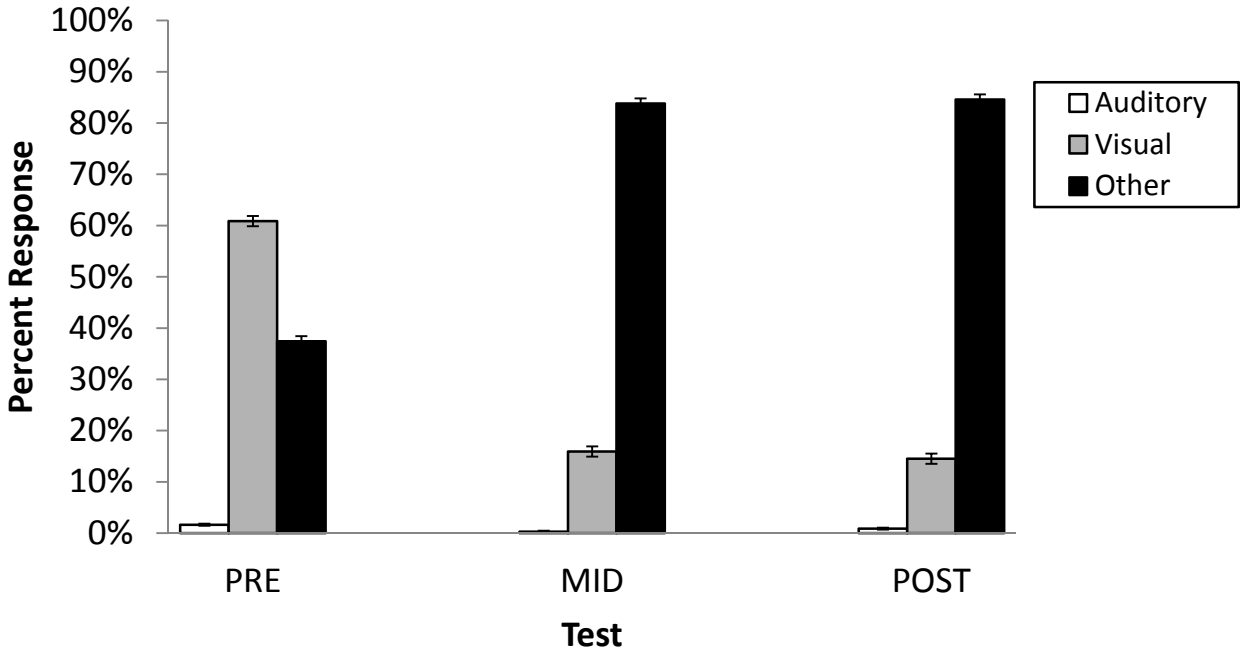

Amount of Integration by Talker
Figure 6
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Amount of Integration
DA
EA
JK
KS
LG
Talker
Pre-test
Post-test


Amount of Integration by Talker (Distefano, 2010)
Figure 7
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Amount of Integration
DA
EA
JK
KS
LG
Talker
Pre-test
Post-test

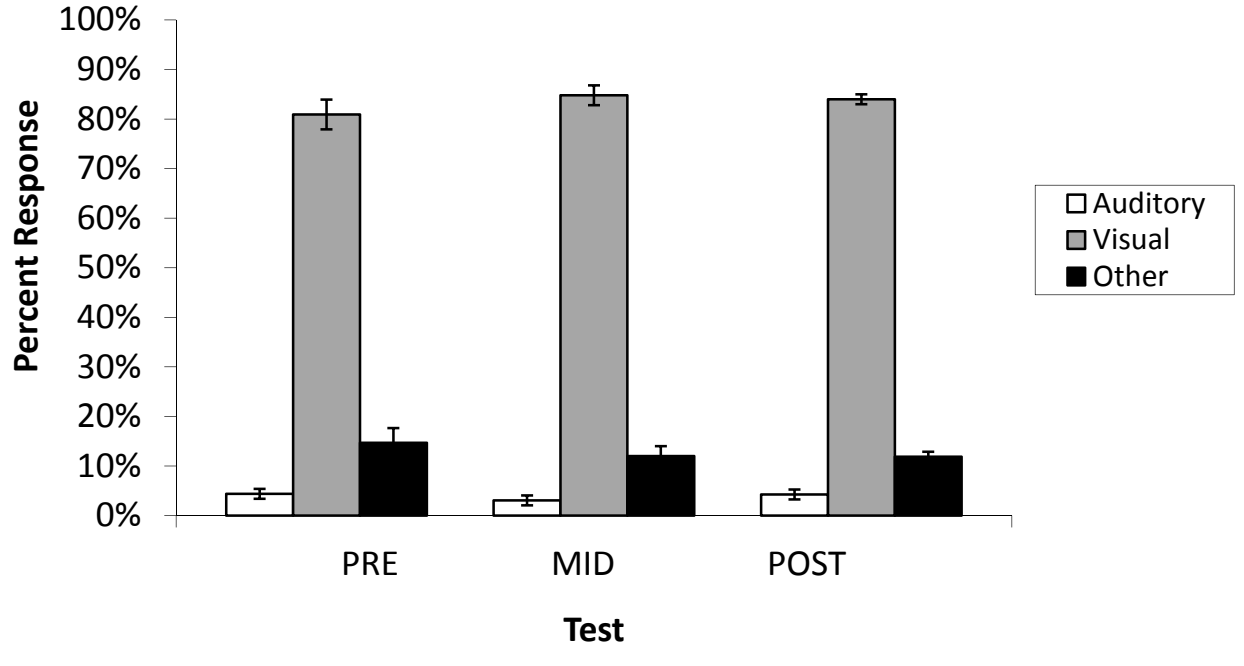

Amount of Integration by Listener
Figure 8
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Amount of Integration
Pre-test
Post-test
NR
AH
HH
MN
TM
Listener


Amount of Integration by Listener (DiStefano, 2010)
Figure 9
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Amount of Integration
Pre-test
Post-test
AF
MO
KD
KH
TP
Listener

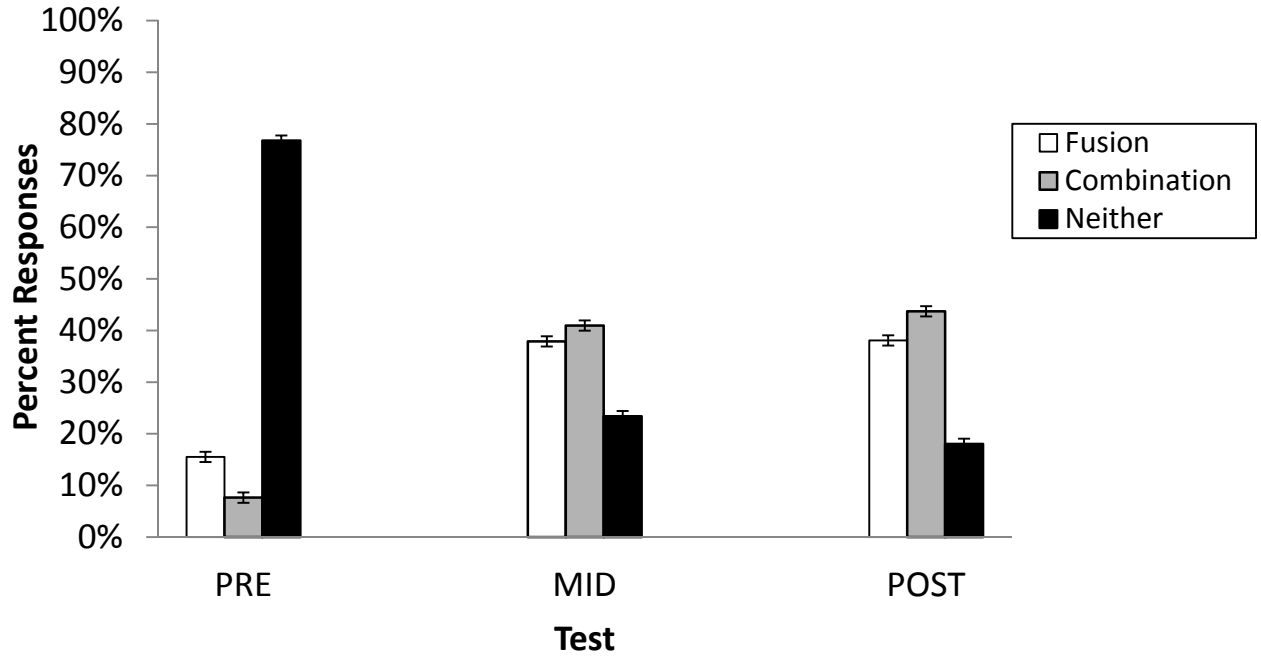
Percent Response Scores Across Tests
Figure 10



Percent Response Scores Across Tests (DiStefano, 2010)
Figure 11



"Other" Responses Across Tests
Figure 12



"Other" Responses Across Tests (DiStefano, 2010)
Figure 13